

# Fixed-Rank Representation for Unsupervised Visual Learning

Risheng Liu, Zhouchen Lin, Fernando De la Torre, and Zhixun Su,

## Abstract

Subspace clustering and feature extraction are two of the most commonly used unsupervised learning techniques in computer vision and pattern recognition. State-of-the-art techniques for subspace clustering make use of recent advances in sparsity and rank minimization. However, existing techniques are computationally expensive and may result in degenerate solutions that degrade clustering performance in the case of insufficient data sampling. To partially solve these problems, and inspired by existing work on matrix factorization, this paper proposes fixed-rank representation (FRR) as a unified framework for unsupervised visual learning. FRR is able to reveal the structure of multiple subspaces in closed-form when the data is noiseless. Furthermore, we prove that under some suitable conditions, even with insufficient observations, FRR can still reveal the true subspace memberships. To achieve robustness to outliers and noise, a sparse regularizer is introduced into the FRR framework. Beyond subspace clustering, FRR can be used for unsupervised feature extraction. As a non-trivial byproduct, a fast numerical solver is developed for FRR. Experimental results on both synthetic data and real applications validate our theoretical analysis and demonstrate the benefits of FRR for unsupervised visual learning.

## Index Terms

Low-Rank Representation, Matrix Factorization, Motion Segmentation, Feature Extraction.

## I. INTRODUCTION

Clustering and embedding are two of the most important techniques for visual data analysis. In the last decade, inspired by the success of compressive sensing, there has been a growing interest in

R. Liu is with School of Mathematical Sciences, Dalian University of Technology, Dalian, P.R. China. This work is done when R. Liu is visiting the Robotics Institute of Carnegie Mellon University. E-mail: rslu0705@gmail.com.

Z. Lin is with Microsoft Research Asia and Key Lab. of Machine Perception (MOE), Peking University, P.R. China.

F. De la Torre is with Robotics Institute, Carnegie Mellon University, USA.

Z. Su is with School of Mathematical Sciences, Dalian University of Technology, Dalian, P.R. China.

incorporating sparsity to visual learning, such as image/video processing [1], object classification [2], [3] and motion segmentation [4]. Early studies [5], [2] usually consider the 1D sparsity (i.e., the nonzero entries of a vector, also known as the  $l_0$  norm) in their models. Recently, there has been a surge of methods [1], [6], [7] which also consider the rank of a matrix as a 2D sparsity measure. However, it is difficult to directly solve these models due to the discrete nature of the  $l_0$  norm and the rank function. A common strategy to alleviate this problem has been to use the  $l_1$  norm and the nuclear norm [8] as the convex surrogates of the  $l_0$  norm and the rank function, respectively.

An important problem in unsupervised learning of visual data is subspace clustering. Recent advances in subspace clustering make use of sparsity-based techniques. For example, sparse subspace clustering (SSC) [5], [9], [10] uses the 1D sparsest representation vectors produced by  $l_1$  norm minimization to define the affinity matrix of an undirected graph. Then subspace clustering is performed by spectral clustering techniques, such as normalized cut (NCut) [11]. However, as SSC computes the sparsest representation of each points individually, there is no global structural constraint on the affinity matrix. This characteristic can degrade the clustering performance when data is grossly corrupted. Moreover, according to the theoretical work of [12], the within subspace connectivity assumption for SSC holds only for 2- and 3-dimensional subspaces. So SSC may probably over-segment subspaces when the dimensions are higher than 3.

Low-rank representation (LRR) [6], [7], [13] is another recently proposed sparsity-based subspace clustering model. The intuition behind LRR is to learn a low-rank representation of the data. The work by [14] shows that LRR is intrinsically equivalent to the shape interaction matrix (SIM) [15] in absence of noise. In this case, LRR can reveal the true clustering when the subspaces are independent and the data sampling is sufficient<sup>1</sup>. However, LRR suffers from some limitations as well. First, the nuclear norm minimization in LRR typically requires to calculate the singular value decomposition (SVD) at each iteration, which becomes computationally impractical as the scale of the problem grows. By combining a linearized version of alternating direction method (ADM) [17] with an acceleration technique for SVD computation, the work in [18] proposed a fast solver, which significantly improves the speed for solving LRR. However, the SVD computation still cannot be completely avoided. Second, and more importantly, if the observations are insufficient, LRR (also SSC) may result in a degenerate solution that significantly degrades the clustering performance. The work in [16] introduces “hidden effects” to overcome this drawback. However, it is unclear whether such “hidden effects” can recover the multiple subspace structure for clustering. Moreover, introducing latent variables makes

<sup>1</sup>The subspaces are independent if and only if the dimension of their direct sum is equal to the sum of their dimensions [14]. For each subspace, the data sampling is sufficient if and only if the rank of the data matrix is equal to the dimension of the subspace [16].

the problem more complex and hard to optimize.

The insufficient data sampling problem in SSC and LRR is similar in spirit to the small sample size problem, that is common in some subspace learning methods, such as linear discriminant analysis [19] and canonical correlation analysis [20]. In these methods, if the number of samples is smaller than the dimension of the features, the covariance matrices are rank deficient. Three are the common approaches to solve this problem [21]: dimensionality reduction, regularization and factorization (i.e., explicitly parameterize the projection matrix as the product of low-rank matrices). In this paper, we incorporate the factorization idea into representation learning and propose fixed-rank representation (FRR) to partially solve the problems in existing unsupervised visual learning models. FRR has three main benefits:

- Unlike SSC and LRR, which use the sparsest and lowest rank representations, FRR *explicitly* parameterizes the representation matrix as the product of two low-rank matrices. When there is no noise and the data sampling is sufficient, we prove that the FRR solution is also the optimal solution to LRR. In this case, FRR can reveal the multiple subspace structure. Furthermore, we prove that under some suitable conditions, even when the data sampling is insufficient, the memberships of samples to each subspace still can be identified by FRR. A sparse regularizer is introduced to FRR to model both small noises and gross outliers, which provides robustness to FRR in real applications.
- The most expensive computational component in LRR is to perform SVD at each iteration. Even with some acceleration techniques, the scalability of the nuclear norm minimization is still limited by the computational complexity of SVD. In contrast, FRR avoids SVD computation and can be efficiently applied to large-scale problems.
- FRR can also be extended for unsupervised feature extraction. By considering a transposed version of FRR (TFRR), we show that FRR is related to existing feature extraction methods, such as principal component analysis (PCA) [22], [23]. Indeed, our analysis provides a unified framework to understand single subspace feature extraction and multiple subspace clustering by analyzing the column and row spaces of the data.

## II. A REVIEW OF PREVIOUS WORK

Given a data set<sup>2</sup>  $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_k] \in \mathbb{R}^{d \times n}$  drawn from a union of  $k$  subspaces  $\{\mathcal{C}_i\}_{i=1}^k$ , where  $\mathbf{X}_i$  is a collection of  $n_i$  data points sampled from the subspace  $\mathcal{C}_i$  with an unknown dimension  $d_{\mathcal{C}_i}$ , the goal of subspace clustering is to cluster data points into their respective subspaces. This section provides a review of SSC and LRR for solving this problem. To clearly understand the mechanism of these methods, we first consider the case when the data is noise-free. From now on, we always write  $\mathbf{X} = \mathbf{U}_X \Sigma_X \mathbf{V}_X^T$  and  $r_X$  as the compact SVD and the rank of  $\mathbf{X}$ , respectively.

### A. Sparse Subspace Clustering (SSC)

SSC [5], [9], [10] is based on the idea that each data point in the subspace  $\mathcal{C}_i$  should be represented as a linear combination of other points that are also in  $\mathcal{C}_i$ . Using this intuition, SSC finds the sparsest representation coefficients  $\mathbf{Z} = [[\mathbf{Z}]_1, [\mathbf{Z}]_2, \dots, [\mathbf{Z}]_n]$  by considering the sequence of optimization problems

$$\min_{[\mathbf{Z}]_i} \|[\mathbf{Z}]_i\|_1, \text{ s.t. } [\mathbf{X}]_i = \mathbf{X}[\mathbf{Z}]_i, [\mathbf{Z}]_{ii} = 0, \quad (1)$$

where  $i = 1, 2, \dots, n$ . Then one can use  $\mathbf{Z}$  to define the affinity matrix of an undirected graph as  $(|\mathbf{Z}| + |\mathbf{Z}^T|)$  and perform NCut on this graph, where  $|\mathbf{Z}|$  denotes a matrix whose entries are the absolute values of  $\mathbf{Z}$ . The SSC model can also be rewritten in matrix form as

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_1, \text{ s.t. } \mathbf{X} = \mathbf{X}\mathbf{Z}, [\mathbf{Z}]_{ii} = 0. \quad (2)$$

Note that both  $l_1$  norm minimization models (1) and (2) can only be solved numerically.

### B. Low-Rank Representation (LRR)

By extending the sparsity measure from 1D to 2D for the representation, LRR [6], [7], [13] proposes a low-rank based criterion for subspace clustering. By utilizing the nuclear norm as a surrogate for the rank function, LRR solves the following nuclear norm minimization problem

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_*, \text{ s.t. } \mathbf{X} = \mathbf{X}\mathbf{Z}. \quad (3)$$

Unlike SSC, which can only be solved numerically,  $\mathbf{V}_X \mathbf{V}_X^T$  (also known as SIM [15]), which has a block-diagonal structure, is the closed-form solution to (3) [14]. Although [14] has proved this, in the following section, we will provide a simpler derivation, that provides new insights into LRR.

<sup>2</sup>Bold capital letters (e.g.,  $\mathbf{M}$ ) denote matrices. The range and the null spaces of  $\mathbf{M}$  are defined as  $\mathcal{R}(\mathbf{M}) := \{\mathbf{a} | \exists \mathbf{b}, \mathbf{a} = \mathbf{M}\mathbf{b}\}$  and  $\mathcal{N}(\mathbf{M}) := \{\mathbf{a} | \mathbf{M}\mathbf{a} = \mathbf{0}\}$ , respectively.  $[\mathbf{M}]_{ij}$  and  $[\mathbf{M}]_i$  denote the  $(i, j)$ -th entry and the  $i$ -th column of  $\mathbf{M}$ , respectively.  $\mathbf{M}^\dagger$  denotes the Moore-Penrose pseudoinverse of  $\mathbf{M}$ . The block-diagonal matrix formed by a collection of matrices  $\mathbf{M}_1, \mathbf{M}_2, \dots, \mathbf{M}_k$  is denoted by  $\text{diag}(\mathbf{M}_1, \mathbf{M}_2, \dots, \mathbf{M}_k)$ .  $\mathbf{1}_n$  is the all-one column vector of length  $n$ .  $\mathbf{I}_n$  is the  $n \times n$  identity matrix.  $\langle \cdot, \cdot \rangle$  denotes the inner product of two matrices. A variety of norms on matrix and vector will be used.  $\|\cdot\|_F$  is the Frobenius norm,  $\|\cdot\|_*$  is the nuclear norm [8],  $\|\cdot\|_{2,1}$  is the  $l_{2,1}$  norm [24],  $\|\cdot\|$  is the spectral norm,  $\|\cdot\|_1$ ,  $\|\cdot\|_2$  and  $\|\cdot\|_\infty$  are the  $l_1$ ,  $l_2$  and  $l_\infty$  norms, respectively.

### III. FIXED-RANK REPRESENTATION

In this section, we propose a new model, named fixed-rank representation (FRR), for subspace clustering. We start with the following analysis on LRR.

#### A. Motivation

To better understand the mechanism of LRR and illustrate our motivation, we show that  $\mathbf{V}_X \mathbf{V}_X^T \in \mathcal{R}(\mathbf{X}^T)$  is the optimal solution to LRR in a simple way<sup>3</sup>. By the identity  $\mathbf{X} = \mathbf{X} \mathbf{X}^\dagger \mathbf{X}$  and the constraint in (3), we have  $\mathbf{X} = \mathbf{X} \mathbf{Z} = \mathbf{X} \mathbf{X}^\dagger \mathbf{X} \mathbf{Z} = \mathbf{X} \mathbf{X}^\dagger \mathbf{X}$ . Thus  $\mathbf{X}^\dagger \mathbf{X} = \mathbf{V}_X \mathbf{V}_X^T$  is a feasible solution to (3). So the general form of the solution is  $\mathbf{Z} = \mathbf{V}_X \mathbf{V}_X^T + \mathbf{Z}_n$ , where  $\mathbf{Z}_n \in \mathcal{N}(\mathbf{X})$ . As  $\mathcal{R}(\mathbf{X}^T) \perp \mathcal{N}(\mathbf{X})$ , we have  $\mathbf{V}_X^T \mathbf{Z}_n = \mathbf{0}$ . This together with the duality definition of nuclear norm [8] leads the following inequality

$$\|\mathbf{Z}\|_* = \max_{\|\mathbf{Y}\| \leq 1} \langle \mathbf{Z}, \mathbf{Y} \rangle \geq \langle \mathbf{Z}, \mathbf{V}_X \mathbf{V}_X^T \rangle = r_X = \|\mathbf{V}_X \mathbf{V}_X^T\|_*.$$

This concludes that  $\mathbf{V}_X \mathbf{V}_X^T$  is the minimizer to (3).

The first observation from the previous analysis is that LRR can successfully remove the effects from  $\mathcal{N}(\mathbf{X})$  to obtain a block-diagonal matrix when the data sampling is sufficient. However, it is also observed that the “lowest rank” representation in LRR is actually the largest rank matrix within the row space of  $\mathbf{X}$ , namely the rank of this representation is always equal to the dimension of the row space. Therefore, the lack of observations for each subspace may significantly degrade the clustering performance. For example, due to insufficient data sampling, the dimension of the row space may be equal to the number of samples (i.e.,  $r_X = n \leq d$ ). In this case, the optimal solution to (3) may reduce to an identity matrix and thus LRR may fail. See Fig. 1 as an example.

An obvious question is whether we can find a lower rank representation in the row space of the data set to exactly reveal the subspace memberships for clustering, even when the data sampling is insufficient. In the following subsection, we give a positive answer to this question.

#### B. The Basic Model

The key idea of FRR is to minimize the Frobenius norm of the representation  $\mathbf{Z}$  instead of the nuclear norm as in LRR. FRR simultaneously computes a fixed lower rank representation  $\tilde{\mathbf{Z}}$  (hereafter we write  $\text{rank}(\tilde{\mathbf{Z}}) = m$ ). That is, we jointly optimize  $\mathbf{Z}$  and  $\tilde{\mathbf{Z}}$  as

$$\min_{\mathbf{Z}, \tilde{\mathbf{Z}}} \|\mathbf{Z} - \tilde{\mathbf{Z}}\|_F^2, \text{ s.t. } \mathbf{X} = \mathbf{X} \mathbf{Z}, \text{ rank}(\tilde{\mathbf{Z}}) = m. \quad (4)$$

<sup>3</sup>Note that here we only analyze the optimality of  $\mathbf{V}_X \mathbf{V}_X^T$  to (3), not its uniqueness.

Obviously,  $\tilde{\mathbf{Z}}$  can be expressed, non-uniquely, as a matrix product  $\tilde{\mathbf{Z}} = \mathbf{L}\mathbf{R}$ , where  $\mathbf{L} \in \mathbb{R}^{n \times m}$  and  $\mathbf{R} \in \mathbb{R}^{m \times n}$ . Replacing  $\tilde{\mathbf{Z}}$  by  $\mathbf{L}\mathbf{R}$ , we arrive at our basic FRR model

$$\min_{\mathbf{Z}, \mathbf{L}, \mathbf{R}} \|\mathbf{Z} - \mathbf{L}\mathbf{R}\|_F^2, \text{ s.t. } \mathbf{X} = \mathbf{X}\mathbf{Z}. \quad (5)$$

In the following sections, we will analyze the problem (5), show properties of the solution to (5), and extend it for real applications.

### C. Analysis on the Basic Model

At first sight, the factorization of  $\tilde{\mathbf{Z}}$  leads to a non-convex optimization problem which may prevent one from getting a global solution. The difficulty results from the fact that the minimizer is non-unique. Fortunately, in the following theorem, we prove that one can always obtain a globally optimal solution to (5) in closed-form.

*Theorem 1:* Let  $[\mathbf{V}_X]_{1:m} = [[\mathbf{V}_X]_1, [\mathbf{V}_X]_2, \dots, [\mathbf{V}_X]_m]$ . Then for any fixed  $m \leq r_X$ ,  $(\mathbf{Z}^*, \mathbf{L}^*, \mathbf{R}^*) := (\mathbf{V}_X \mathbf{V}_X^T, [\mathbf{V}_X]_{1:m}, [\mathbf{V}_X]_{1:m}^T)$  is a globally optimal solution to (5) and the minimum objective function value is  $(r_X - m)$ .

The proof of this theorem is based on the following lemma.

*Lemma 2:* (Courant-Fischer Minimax Theorem [25]) For any symmetric matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$ , we have that

$$\lambda_i(\mathbf{A}) = \max_{\dim(\mathcal{S})=i} \min_{\mathbf{0} \neq \mathbf{y} \in \mathcal{S}} \mathbf{y}^T \mathbf{A} \mathbf{y} / \mathbf{y}^T \mathbf{y}, \text{ for } i = 1, 2, \dots, n,$$

where  $\mathcal{S} \subset \mathbb{R}^n$  is some subspace and  $\lambda_i(\mathbf{A})$  is the  $i$ -th largest eigenvalue of  $\mathbf{A}$ .

*Proof:* First, by the well known Eckart-Young theorem [26], given  $\mathbf{Z}$ , we have

$$\min_{\mathbf{L}, \mathbf{R}} \|\mathbf{Z} - \mathbf{L}\mathbf{R}\|_F^2 = \sum_{i=m+1}^d \sigma_i^2(\mathbf{Z}), \quad (6)$$

where  $\sigma_i(\mathbf{Z})$  is the  $i$ -th largest singular value of  $\mathbf{Z}$ . Now we prove that

$$\text{if } \mathbf{X} = \mathbf{X}\mathbf{Z} \text{ then } \sigma_{r_X}(\mathbf{Z}) \geq 1. \quad (7)$$

By  $\mathbf{X} = \mathbf{X}\mathbf{Z}$ , we have that  $\text{rank}(\mathbf{Z}) \geq r_X$ . Then (6) and (7) imply that the minimum objective function value is no less than  $r_X - m$ . Indeed, by the compact SVD of  $\mathbf{X}$  and  $\mathbf{X} = \mathbf{X}\mathbf{Z}$ , we have

$$\mathbf{V}_X^T = \mathbf{V}_X^T \mathbf{Z}, \quad (8)$$

By Lemma 2,  $\sigma_i(\mathbf{Z}) = \max_{\dim(\mathcal{S})=i} \min_{\mathbf{0} \neq \mathbf{y} \in \mathcal{S}} \|\mathbf{Z}^T \mathbf{y}\|_2 / \|\mathbf{y}\|_2$ , where  $\|\cdot\|_2$  is the  $l_2$  norm of a vector. So by choosing  $\mathcal{S} = \mathcal{R}(\mathbf{V}_X)$  and utilizing (8),

$$\begin{aligned} \sigma_{r_X}(\mathbf{Z}) &\geq \min_{\mathbf{0} \neq \mathbf{y} \in \mathcal{R}(\mathbf{V}_X)} \|\mathbf{Z}^T \mathbf{y}\|_2 / \|\mathbf{y}\|_2 \\ &= \min_{\mathbf{b} \neq \mathbf{0}} \|\mathbf{Z}^T \mathbf{V}_X \mathbf{b}\|_2 / \|\mathbf{V}_X \mathbf{b}\|_2 \\ &= \min_{\mathbf{b} \neq \mathbf{0}} \|\mathbf{V}_X \mathbf{b}\|_2 / \|\mathbf{V}_X \mathbf{b}\|_2 = 1. \end{aligned} \quad (9)$$

Next, when  $\mathbf{Z} = \mathbf{V}_X \mathbf{V}_X^T$ , it can be easily checked that the objective function value is  $(r_X - m)$ . Again, by Eckart-Young theorem,  $\mathbf{L}\mathbf{R} = [\mathbf{V}_X]_{1:m} [\mathbf{V}_X]_{1:m}^T$ . Thus we have  $(\mathbf{V}_X \mathbf{V}_X^T, [\mathbf{V}_X]_{1:m}, [\mathbf{V}_X]_{1:m}^T)$  is a globally optimal solution to (5), thereby completing the proof of the theorem. ■

Based on Theorem 1, we can derive the following corollary to illustrate the structure of the optimal solution to (5).

*Corollary 3:* Under the assumption that subspaces are independent and data  $\mathbf{X}$  is clean, there exists a globally optimal solution  $(\mathbf{Z}^*, \mathbf{L}^*, \mathbf{R}^*)$  to problem (5) with the following structure:

$$\mathbf{Z}^* = \text{diag}(\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_k), \quad (10)$$

where  $\mathbf{Z}_i$  is an  $n_i \times n_i$  matrix with  $\text{rank}(\mathbf{Z}_i) = d_{C_i}$  and

$$\mathbf{L}^* \mathbf{R}^* \in \mathcal{R}(\mathbf{Z}^*) = \mathcal{R}(\mathbf{X}^T). \quad (11)$$

The proof of this corollary is based on the following lemma.

*Lemma 4:* [15] Let  $\mathbf{X} = \mathbf{U}_X \Sigma_X \mathbf{V}_X^T$  be the compact SVD. Under the same assumption in Corollary 3,  $\mathbf{V}_X \mathbf{V}_X^T$  is a block diagonal matrix that has exactly  $k$  blocks. Moreover, the  $i$ -th block on its diagonal is an  $n_i \times n_i$  matrix with rank  $d_{C_i}$ .

*Proof:* By the proof of Theorem 1, we have that  $\mathbf{Z}^* = \mathbf{V}_X \mathbf{V}_X^T$  is a global optimal solution to (5) and any global optimal  $\mathbf{L}^*$  and  $\mathbf{R}^*$  are in the range space  $\mathcal{R}(\mathbf{Z}^*)$ . So we have that  $\mathbf{L}^* \mathbf{R}^* \in \mathcal{R}(\mathbf{Z}^*) = \mathcal{R}(\mathbf{X}^T)$ . By Lemma 4, we achieve the block diagonal structure (10) for  $\mathbf{Z}^*$ , which concludes the proof. ■

However, such  $\mathbf{Z}^*$  suffers from the same limitation of LRR. Namely, when the data sampling is insufficient,  $\mathbf{Z}^*$  will probably degenerate and thus the clustering may fail.

Fortunately, as shown in (11),  $\mathbf{L}^* \mathbf{R}^*$  can still be spanned by the row space of  $\mathbf{X}$ . This inspires us to consider this lower rank representation for subspace clustering.

*Corollary 5:* Assuming that the columns of  $\mathbf{Z}^*$  are normalized (i.e.  $\mathbf{1}_n^T \mathbf{Z}^* = \mathbf{1}_n^T$ ) and fix  $m = k$ , then there exists globally optimal  $\mathbf{L}^*$  and  $\mathbf{R}^*$  to problem (5) such that

$$\mathbf{L}^* \mathbf{R}^* = \text{diag}(n_1 \mathbf{1}_{n_1} \mathbf{1}_{n_1}^T, n_2 \mathbf{1}_{n_2} \mathbf{1}_{n_2}^T, \dots, n_k \mathbf{1}_{n_k} \mathbf{1}_{n_k}^T). \quad (12)$$

**Remark:** Corollary 5 does not guarantee that an arbitrary rank- $k$  optimal solution has the block-diagonal structure (12) due to the non-unique of the minimizer  $(\mathbf{L}^*, \mathbf{R}^*)$ . However, in our experiments, we have observed that empirically choosing the first  $k$  columns of  $\mathbf{V}_X$  works well on the tested data (e.g., Fig. 1).

*Proof:* By Corollary 3 and the normalization assumption,  $\mathbf{Z}^* = \text{diag}(\mathbf{Z}_1^*, \mathbf{Z}_2^*, \dots, \mathbf{Z}_k^*)$ , where  $\mathbf{Z}_i^*$  is an  $n_i \times n_i$  for subspace  $\mathcal{C}_i$  and  $\mathbf{1}_{n_i}$  is an eigenvector of  $\mathbf{Z}_i^*$  with eigenvalue 1. Thus there exists a basis  $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_k]$ , each vector of which with the form  $\mathbf{h}_i = [\mathbf{0}, \mathbf{1}_{n_i}^T, \mathbf{0}]^T$  is eigenvector of  $\mathbf{Z}$

with eigenvalue 1. By the Eckart-Young theorem (similar to the proof of Theorem 1), we have that  $\mathbf{L}^* = \mathbf{H}$  and  $\mathbf{R}^* = \mathbf{H}^T$  are global optimal solutions to (5), which directly leads (12). ■

In principle, the normalization of  $\mathbf{Z}^*$  could be considered as a strong assumption, hence it cannot always be guaranteed in real situations. Therefore, we explicitly enforce each column of  $\mathbf{Z}$  to sum to one

$$\min_{\mathbf{Z}, \mathbf{L}, \mathbf{R}} \|\mathbf{Z} - \mathbf{L}\mathbf{R}\|_F^2, \text{ s.t. } \mathbf{X} = \mathbf{X}\mathbf{Z}, \mathbf{1}_n^T \mathbf{Z} = \mathbf{1}_n^T. \quad (13)$$

#### D. Sparse Regularization for Corruptions

In real applications, the data are often corrupted by both small noises and gross outliers. In the following, we show how to extend problem (13) to deal with corruptions. By modeling corruptions as a new term  $\mathbf{E}$ , we consider the following regularized optimization problem

$$\begin{aligned} \min_{\mathbf{Z}, \mathbf{L}, \mathbf{R}, \mathbf{E}} \quad & \|\mathbf{Z} - \mathbf{L}\mathbf{R}\|_F^2 + \mu \|\mathbf{E}\|_s, \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{E}, \mathbf{1}_n^T \mathbf{Z} = \mathbf{1}_n^T, \end{aligned} \quad (14)$$

where the parameter  $\mu > 0$  is used to balance the effects of the two terms and  $\|\cdot\|_s$  is a sparse norm corresponding to our assumption on  $\mathbf{E}$ . Here we adopt the  $l_{2,1}$  norm to characterize the corruptions since it can successfully identify the indices of the outliers and remove small noises [27]. Algorithm 1 summarizes the whole FRR based subspace clustering framework.

---

#### Algorithm 1 FRR for Subspace Clustering

---

**Input:** Let  $\mathbf{X} \in \mathbb{R}^{d \times n}$  be a set of data points sampled from  $k$  subspaces.

**Step 1:** Solve (14) to obtain  $(\mathbf{Z}^*, \mathbf{L}^*, \mathbf{R}^*)$ .

**Step 2:** Construct a graph by using  $(|\mathbf{Z}^*| + |(\mathbf{Z}^*)^T|)$  or  $(|\mathbf{L}^* \mathbf{R}^*| + |(\mathbf{L}^* \mathbf{R}^*)^T|)$  as the affinity matrix.

**Step 3:** Apply NCut to this graph to obtain the clustering.

---

### IV. EXTENDING FRR FOR FEATURE EXTRACTION

Besides subspace clustering, the mechanism of FRR can also be applied for feature extraction. That is, one can recover the column space of the data set by solving the following transposed FRR (TFRR)

$$\min_{\mathbf{Z}, \mathbf{L}, \mathbf{R}} \|\mathbf{Z} - \mathbf{L}\mathbf{R}\|_F^2, \text{ s.t. } \mathbf{X} = \mathbf{Z}\mathbf{X}, \quad (15)$$

where  $m \leq r_X$ ,  $\mathbf{L} \in \mathbb{R}^{d \times m}$ ,  $\mathbf{R} \in \mathbb{R}^{m \times d}$  and  $\mathbf{Z} \in \mathbb{R}^{d \times d}$ . For noisy data, by using similar techniques as in Section III-D, we introduce an explicit corruption term  $\mathbf{E}$  into the objective function and the constraint. Hence we obtain the robust version of TFRR for feature extraction

$$\min_{\mathbf{Z}, \mathbf{L}, \mathbf{R}, \mathbf{E}} \|\mathbf{Z} - \mathbf{L}\mathbf{R}\|_F^2 + \mu \|\mathbf{E}\|_s, \text{ s.t. } \mathbf{X} = \mathbf{Z}\mathbf{X} + \mathbf{E}. \quad (16)$$

### A. Relationship to Principal Component Analysis

Principal component analysis (PCA) is one of the most popular dimensionality reduction techniques [22], [23]. The basic ideas behind PCA date back to Pearson in 1901 [22], and a more general procedure was described by Hotelling [23] in 1933. There are several energy functions which lead to subspace spanned by the principal components [21]. For instance, PCA finds the matrix  $\mathbf{P} \in \mathbb{R}^{d \times m}$  that minimizes:

$$\min_{\mathbf{P}} \|\mathbf{X} - \mathbf{P}\mathbf{P}^T\mathbf{X}\|_F^2, \text{ s.t. } \mathbf{P}^T\mathbf{P} = \mathbf{I}_m. \quad (17)$$

It can be shown that  $\mathbf{P}^* = [\mathbf{U}_X]_{1:m}$  is the optimal solution to (17), where  $[\mathbf{U}_X]_{1:m} = [[\mathbf{U}_X]_1, [\mathbf{U}_X]_2, \dots, [\mathbf{U}_X]_m]$ . The following corollary shows that the mechanism of TFRR can also be applied to formulate PCA.

*Corollary 6:* For any fixed  $m \leq r_X$ ,  $(\mathbf{Z}^*, \mathbf{L}^*, \mathbf{R}^*) := (\mathbf{U}_X \mathbf{U}_X^T, [\mathbf{U}_X]_{1:m}, [\mathbf{U}_X]_{1:m}^T)$  is a globally optimal solution to (15) and the minimum objective function value is  $(r_X - m)$ .

*Proof:* The proof of Theorem 1 directly leads to the above corollary. ■

## V. OPTIMIZATION FOR FRR

In this section, we develop a fast numerical solver for FRR related models by extending the classic alternating direction method (ADM) [17] to non-convex problems. To solve the problem (14)<sup>4</sup>, we introduce Lagrange multipliers  $\Lambda$  and  $\Pi$  to remove the equality constraints. The resulting augmented Lagrangian function is

$$\begin{aligned} \mathcal{L}_A(\mathbf{Z}, \mathbf{L}, \mathbf{R}, \mathbf{E}, \Lambda, \Pi) &= \|\mathbf{Z} - \mathbf{L}\mathbf{R}\|_F^2 + \mu \|\mathbf{E}\|_{2,1} \\ &+ \langle \Lambda, \mathbf{X} - \mathbf{X}\mathbf{Z} - \mathbf{E} \rangle + \langle \Pi, \mathbf{1}_n^T \mathbf{Z} - \mathbf{1}^T \rangle \\ &+ \frac{\beta}{2} (\|\mathbf{X} - \mathbf{X}\mathbf{Z} - \mathbf{E}\|_F^2 + \|\mathbf{1}_n^T \mathbf{Z} - \mathbf{1}_n\|_F^2), \end{aligned} \quad (18)$$

where  $\beta > 0$  is a penalty parameter. It is important to note that although (18) is not jointly convex for all variables, it is convex with respect to each variable while fixing the others. This property allows the iteration scheme to be well defined. So we minimize (18) with respect to  $\mathbf{L}$ ,  $\mathbf{R}$ ,  $\mathbf{Z}$ , and  $\mathbf{E}$  one at a time while fixing the others at their latest values, and then update the Lagrange multipliers  $\Lambda$  and

<sup>4</sup>As other FRR related models can be solved in similar way, we do not further explore them in this section.

II:

$$\mathbf{L}_+ \leftarrow \mathbf{Z}\mathbf{R}^\dagger \equiv \mathbf{Z}\mathbf{R}^T(\mathbf{R}\mathbf{R}^T)^\dagger, \quad (19)$$

$$\mathbf{R}_+ \leftarrow \mathbf{L}_+^\dagger \mathbf{Z} \equiv (\mathbf{L}_+^T \mathbf{L}_+)^{\dagger} \mathbf{L}_+^T \mathbf{Z}, \quad (20)$$

$$\mathbf{Z}_+ \leftarrow (2\mathbf{I}_n + \beta(\mathbf{X}^T \mathbf{X} + \mathbf{1}_n \mathbf{1}_n^T))^{-1} \mathbf{B}, \quad (21)$$

$$\mathbf{E}_+ \leftarrow \arg \min_{\mathbf{E}} \mu \|\mathbf{E}\|_{2,1} + \frac{\beta}{2} \|\mathbf{C} - \mathbf{E}\|_F^2, \quad (22)$$

$$\Lambda_+ \leftarrow \Lambda + \beta(\mathbf{X} - \mathbf{X}\mathbf{Z}_+ - \mathbf{E}_+), \quad (23)$$

$$\Pi_+ \leftarrow \Pi + \beta(\mathbf{1}_n^T \mathbf{Z}_+ - \mathbf{1}_n^T), \quad (24)$$

$$\beta_+ \leftarrow \min(\bar{\beta}, \rho\beta), \quad (25)$$

where the subscript  $+$  denotes that the values are updated,  $\bar{\beta}$  is the upper bound of  $\beta$ ,  $\rho > 1$  is the step length parameter,  $\mathbf{B} = 2\mathbf{L}_+ \mathbf{R}_+ + \beta(\mathbf{X}^T \mathbf{X} - \mathbf{X}^T(\mathbf{E} - \Lambda/\beta)) + \beta \mathbf{1}_n \mathbf{1}_n^T - \mathbf{1}_n \Pi$  and  $\mathbf{C} = \mathbf{X} - \mathbf{X}\mathbf{Z}_+ + \Lambda/\beta$ . The subproblem (22) can be solved by Lemma 3.2 in [6]. We then reduce the computational cost for solving (19) and (20). It follows from (20) that

$$\mathbf{L}_+ \mathbf{R}_+ = \mathbf{L}_+ (\mathbf{L}_+^T \mathbf{L}_+)^{\dagger} \mathbf{L}_+^T \mathbf{Z} = \mathcal{P}_{L_+}(\mathbf{Z}). \quad (26)$$

By considering the compact SVD:  $\mathbf{R} = \mathbf{U}_{R_r} \Sigma_{R_r} \mathbf{V}_{R_r}^T$ , we have  $\mathbf{L}_+ = \mathbf{Z} \mathbf{V}_{R_r} \Sigma_{R_r}^{-1} \mathbf{U}_{R_r}^T$  and  $\mathbf{Z} \mathbf{R}^T = \mathbf{Z} \mathbf{V}_{R_r} \Sigma_{R_r} \mathbf{U}_{R_r}^T$ . This implies that  $\mathcal{R}(\mathbf{L}_+) = \mathcal{R}(\mathbf{Z} \mathbf{R}^T) = \mathcal{R}(\mathbf{Z} \mathbf{V}_{R_r})$  and

$$\mathbf{L}_+ \mathbf{R}_+ = \mathcal{P}_{Z \mathbf{R}^T}(\mathbf{Z}), \quad (27)$$

where  $\mathcal{P}_{Z \mathbf{R}^T}$  is the orthogonal projection into  $\mathcal{R}(\mathbf{Z} \mathbf{R}^T)$ . Since the objective function of (14) depends on the product  $\mathbf{L}_+ \mathbf{R}_+$ , different values of  $\mathbf{L}_+$  and  $\mathbf{R}_+$  are essentially equivalent as long as they give the same product. The identity (27) shows that the inversion  $(\mathbf{R} \mathbf{R}^T)^\dagger$  and  $(\mathbf{L}_+^T \mathbf{L}_+)^{\dagger}$  can be saved when the projection  $\mathcal{P}_{Z \mathbf{R}^T}$  is computed. Specifically, one can compute  $\mathcal{P}_{Z \mathbf{R}^T} = \mathbf{Q} \mathbf{Q}^T$ , where  $\mathbf{Q}$  is the QR factorization of  $\mathbf{Z} \mathbf{R}^T$ . Then we have  $\mathbf{L}_+ \mathbf{R}_+ = \mathbf{Q} \mathbf{Q}^T \mathbf{Z}$  and one can derive:

$$\mathbf{L}_+ \leftarrow \mathbf{Q}, \quad (28)$$

$$\mathbf{R}_+ \leftarrow \mathbf{Q}^T \mathbf{Z}. \quad (29)$$

The schemes (28) and (29) are often preferred since computing (29) by QR factorization is generally more stable than solving the normal equations [28]. The complete algorithm is summarized in Algorithm 2.

## VI. EXPERIMENTAL RESULTS

This section compared the performance of FRR against state-of-the-art algorithms on both subspace clustering and feature extraction. All experiments are performed on a notebook computer with an Intel Core i7 CPU at 2.00 GHz and 6GB of memory, running Windows 7 and Matlab version 7.10.

**Algorithm 2** Solving (14) by ADM-type Algorithm

---

**Input:** Observation matrix  $\mathbf{X} \in \mathbb{R}^{d \times n}$ ,  $m > 0$ ,  $\epsilon_1, \epsilon_2 > 0$ , parameters  $\beta > 0$  and  $\rho > 1$ .

**Initialization:** Initialize  $\mathbf{Z}_0 \in \mathbb{R}^{n \times n}$ ,  $\mathbf{L}_0 \in \mathbb{R}^{n \times m}$ ,  $\mathbf{R}_0 \in \mathbb{R}^{m \times n}$ ,  $\mathbf{E}_0 \in \mathbb{R}^{d \times n}$ ,  $\Lambda_0 \in \mathbb{R}^{d \times n}$  and  $\Pi_0 \in \mathbb{R}^{1 \times n}$ .

**while** not converged **do**

**Step 1:** Update  $(\mathbf{Z}, \mathbf{L}, \mathbf{R}, \mathbf{E}, \Lambda, \Pi)$  by (28), (29) and (21)–(25).

**Step 2:** Check the convergence conditions:

$$\|\mathbf{X} - \mathbf{X}\mathbf{Z}_+ - \mathbf{E}_+\|_\infty \leq \epsilon_1 \text{ and } \|\mathbf{1}_n^T \mathbf{Z}_+ - \mathbf{1}_n^T\|_\infty \leq \epsilon_2.$$

**end while**

**Output:**  $\mathbf{Z}^*$ ,  $\mathbf{L}^*$ ,  $\mathbf{R}^*$  and  $\mathbf{E}^*$ .

---

*A. Subspace Clustering*

We first consider the subspace clustering problem, and compare the clustering performance and computational speed of FRR to existing state-of-the-art methods, such as SIM, Random Sample Consensus (RANSAC) [29], Local Subspace Analysis (LSA) [30], SSC and LRR. As shown in Section III, both  $\mathbf{Z}$  and  $\mathbf{LR}$  can be utilized for clustering, we call these two strategies  $\text{FRR}_1$  and  $\text{FRR}_2$ , respectively.

1) *Synthetic Data:* We performed subspace clustering on synthetic data to illustrate the insufficient data sampling problem (to verify the analysis in Section III). Let  $k$ ,  $p$ ,  $d_h$  and  $d_l$  denote the number of subspaces, the number of points in each subspace, the features (i.e., observed dimension) and the intrinsic dimension of the subspace, respectively. Then the data set, parameterized as  $(k, p, d_h, d_l)$ , is generated by the same procedure in [6]:  $k$  independent subspaces  $\{\mathcal{C}_i\}_{i=1}^k$  are constructed, whose basis  $\{\mathbf{U}_i\}_{i=1}^k$  are computed by  $\mathbf{U}_{i+1} = \mathbf{T}\mathbf{U}_i$ ,  $1 \leq i \leq k-1$ , where  $\mathbf{T}$  is a random rotation and  $\mathbf{U}_1$  is a random column orthogonal matrix of dimension  $d_h \times d_l$ . Then we construct a  $d_h \times kp$  data matrix  $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_k]$  by sampling  $p$  data vectors from each subspace by  $\mathbf{X}_i = \mathbf{U}_i \mathbf{C}_i$ ,  $1 \leq i \leq k$ , with  $\mathbf{C}_i$  being a  $d_l \times p$  matrix with uniform distribution. To generate the point set for insufficient data sampling clustering, we fix  $k = 10$ ,  $d_h = 100$  and  $d_l = 50$  and vary  $p \in [10, 30]$ . In this way, the number of samples in each subspace (at most 30) is less than the intrinsic dimension (50 for each subspace).

Fig. 1 illustrated the structures of  $\mathbf{Z} = \mathbf{V}_X \mathbf{V}_X^T$  and  $\mathbf{LR} = [\mathbf{V}_X]_{1:k} [\mathbf{V}_X]_{1:k}^T$  when  $p = 10$ . Since the data sampling is insufficient, the optimal  $\mathbf{Z}$  for (3) and (5) reduces to  $\mathbf{I}_n$  (see Fig. 1 (a)). In contrast,  $\mathbf{LR}$  can successfully reveal the multiple subspace structure (see Fig. 1 (b)).

We also compared the clustering performances of  $\mathbf{Z}$  and  $\mathbf{LR}$  on the generated data. Fig. 2 shows the clustering accuracy as a function of the number of points. It can be seen that the clustering

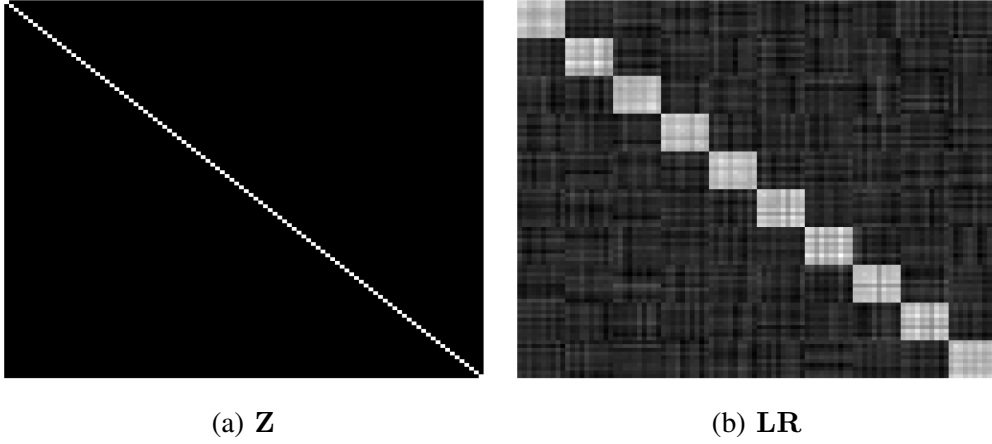


Fig. 1. The structures of  $\mathbf{Z}$  and  $\mathbf{LR}$ , where  $\text{rank}(\mathbf{Z}) = \text{rank}(\mathbf{X}) = kp = 100$  and  $\text{rank}(\mathbf{LR}) = k = 10$ , respectively.

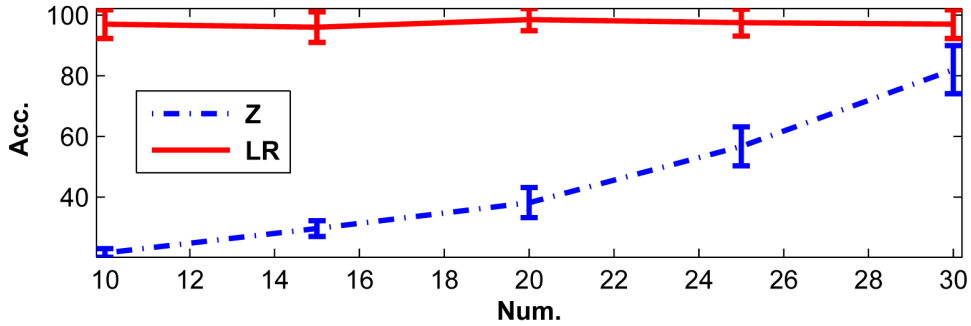


Fig. 2. The mean and std. clustering accuracies (%) of  $\mathbf{Z}$  and  $\mathbf{LR}$  over 20 runs. The  $x$ -axis represents the number of samples in each subspace and the  $y$ -axis represents the clustering accuracy.

accuracy of  $\mathbf{Z}$  is very sensitive to the particular sampling. Although it performs better when  $p$  is increasing, the highest clustering accuracy is only around 80% ( $p = 30$ ). In contrast,  $\mathbf{LR}$  achieves almost perfect results on all data sets. This confirms that the affinity matrix calculated from  $\mathbf{LR}$  can successfully overcome the drawback of using  $\mathbf{Z}$  in (5) and LRR (also SIM) when the data sampling is insufficient.

2) *Motion Segmentation*: Motion segmentation refers to the problem of segmenting tracked feature point trajectories of multiple moving objects in a video sequence. As shown in [4], all the tracked points from a single rigid motion lie in a four-dimensional linear subspace. So this task can be regarded as a subspace clustering problem. We perform the experiments on the Hopkins155 database [31], which is an extensive benchmark for motion segmentation. This database consists of 156 sequences of two or three motions thus there are 156 clustering tasks in total. For a fair comparison, we apply all algorithms to the raw data and the parameters of these methods have been tuned to the best.

We reported the segmentation errors in Table III and presented the percentage of sequences for which the segmentation error is less than or equal to a given percentage of misclassification in Fig. 3.

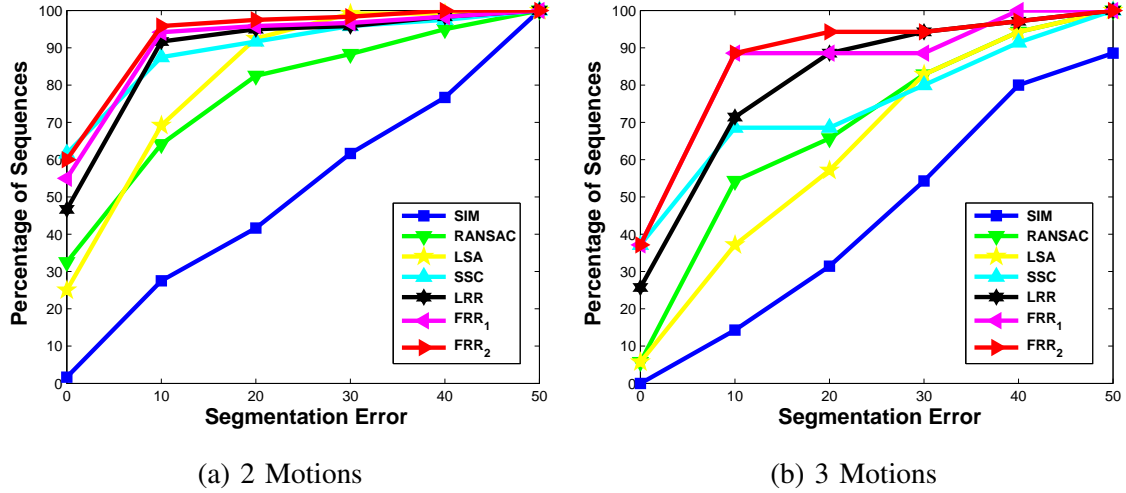


Fig. 3. Percentage of sequences for which the segmentation error is less than or equal to a given percentage of misclassification.

TABLE I  
SEGMENTATION ERRORS (%) ON HOPKINS155 RAW DATA.

Method	2 Motions				3 Motions				All (156)			
	mean	median	std.	max.	mean	median	std.	max.	mean	median	std.	max.
SIM	24.1	24.8	15.4	49.2	27.9	28.5	15.8	64.1	25.1	25.3	15.7	64.1
RANSAC	9.6	3.3	13.1	49.3	13.8	7.8	13.7	44.7	10.8	4.2	13.5	49.3
LSA	6.8	2.8	8.0	40.9	16.8	15.6	12.6	46.6	9.1	4.8	10.1	46.6
SSC	3.7	<b>0.0</b>	9.7	49.9	11.4	3.3	15.0	44.6	5.5	<b>0.0</b>	11.6	49.9
LRR	3.2	0.3	8.2	40.3	7.8	2.8	10.3	41.5	4.3	0.6	8.9	<b>41.5</b>
FRR <sub>1</sub>	2.5	<b>0.0</b>	7.4	40.8	5.9	1.4	10.9	<b>39.4</b>	3.5	<b>0.0</b>	8.9	41.8
FRR <sub>2</sub>	<b>1.8</b>	<b>0.0</b>	<b>5.3</b>	<b>36.1</b>	<b>4.7</b>	<b>1.0</b>	<b>9.1</b>	41.5	<b>2.6</b>	<b>0.0</b>	<b>6.5</b>	<b>41.5</b>

It can be noticed that the performances of three sparsity-based models (i.e., SSC, LRR and FRR) are better than other methods. SSC is worse than LRR because the 1D  $l_1$  norm based criterion finds the representation coefficients of each vector individually, and there is no global constrain. Although the basic forms of LRR (3) and FRR (5) share the same optimal solution to  $\mathbf{Z}$ , FRR<sub>1</sub> performs even better than LRR in real data set. This is because enforcing the normalization constraint in (14) can improve the performance for clustering. Overall, FRR<sub>2</sub> outperforms all other methods in this paper. This result, again, confirms that  $\mathbf{LR}$  in FRR<sub>2</sub> is better than the general  $\mathbf{Z}$  in LRR and FRR<sub>1</sub> for subspace clustering.

For three sparsity-based methods, Table II reports the time in seconds. We can see that the computational time of SSC is lower than the standard LRR. This is because the  $l_1$  norm minimizations in SSC can be solved in parallel and there is only a thresholding process needed at each iteration. While LRR is solved with an SVD in each iteration, and it does not scale well with large number of

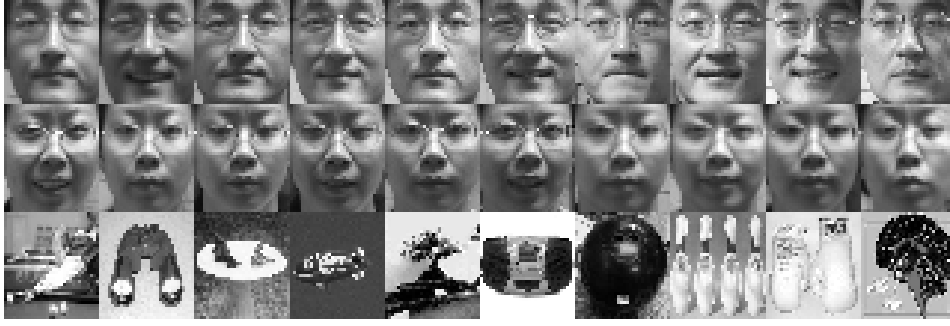


Fig. 4. Examples of the FRGC-Caltech data set. The top two rows correspond to face images and the bottom row non-face images.

samples. By combining linearized ADM with an acceleration technique for SVD, the work in [18] proposed a fast solver for LRR. The running time of this approach is even less than SSC. Our FRR, again, achieves the highest efficiency because it completely avoids SVD computation in the iterations.

TABLE II

THE AVERAGE RUNNING TIME (SECONDS) PER SEQUENCE FOR THREE SPARSITY-BASED METHODS. LRR(A) DENOTES THE ACCELERATED LRR PROPOSED IN [18].

Method	2 Motions	3 Motions	All (156)
SSC	3.5445	7.8493	4.5057
LRR	38.5156	115.3140	55.6259
LRR(A)	1.9415	3.6788	2.3319
FRR	<b>0.9990</b>	<b>2.2799</b>	<b>1.2847</b>

### B. Feature Extraction and Outlier Detection

This experiment tested the effectiveness of TFRR for feature extraction in presence of occlusions. To simulate sample-outliers, we created a dataset by combining images with faces from the FRGC version 2 [32] and images non containing faces from Caltech-256 [33]. We selected 20 images for the first 180 subjects of the FRGC database, having a total of 3600 images. For Caltech-256 database, which contains 257 image categories, we randomly selected 1 image from each class (a total of 257 non-facial images). All images are resized to  $32 \times 36$  and the pixel values are normalized to  $[0, 1]$ . As shown in Fig. 4, there are two types of corruptions: small errors in the facial images (e.g., illuminations and occlusions) and non-facial outliers.

The goal of this task is to robustly extract facial features and use them for classification. That is, we learn a mapping  $\mathbf{P}$  between high dimensional observations and low dimensional features using TFRR,

and identify outliers in the training set by  $\mathbf{E}$ . Then for a new testing data  $\mathbf{x}$ , the feature vector  $\mathbf{y}$  can be computed as  $\mathbf{y} = \mathbf{P}\mathbf{x}$ . We selected the first  $k$  ( $k = 40, 80$ ) identities and 257 non-facial images as the training set and the remaining  $(180 - k)$  identities of facial images for test. We compared two TFRR based strategies (one is directly using  $\mathbf{P} = \mathbf{Z}$ , called TFRR<sub>1</sub>, and another is computing the orthogonal basis  $\mathbf{P} = \text{orth}(\mathbf{L}\mathbf{R})$ , called TFRR<sub>2</sub>) with the “Raw data” baseline and other state-of-the-art approaches, such as PCA, Locality Preserving Projection (LPP) [34] and Neighborhood Preserving Embedding (NPE) [35]. The parameters and the feature dimensions of all methods are tuned to the best for each training set. Table III demonstrates that the performances of TFRR<sub>1</sub> and TFRR<sub>2</sub> are both significantly better than the baseline and PCA. Moreover, TFRR<sub>2</sub> outperforms all other methods on these experiments.

As shown in Fig. 5, the main advantage of TFRR based methods comes from their ability of extracting intrinsic facial features and removing outliers. One can see that most of the intrinsic facial features can be projected into the range space (modeled by  $\mathbf{Z}\mathbf{X}$ , see the middle row), while the small errors of the facial images (e.g., illuminations and occlusions) and non-facial outliers (modeled by  $\mathbf{E}$ ) can be automatically removed (see the bottom row).

TABLE III

CLASSIFICATION ACCURACIES (MEAN  $\pm$  STD.%) ON FRGC-CALTECH DATA SET. “Gm/Pn” MEANS IN THE TESTING DATA  $m$  IMAGES OF EACH SUBJECT ARE RANDOMLY SELECTED AS GALLERY SET AND THE REMAINING  $n$  IMAGES AS PROBE SET. SUCH A TRIAL IS REPEATED 20 TIMES. THE FEATURE DIMENSIONS ARE: PCA (410D, 358D), LPP (170D, 200D), NPE(320D, 160D) AND TFRR<sub>2</sub> (190D, 100D). THE DIMENSION OF THE FEATURE VECTOR PRODUCED BY TFRR<sub>1</sub> IS THE SAME AS THE OBSERVED DATA.

Train	Test	Raw	PCA	LPP	NPE	TFRR <sub>1</sub>	TFRR <sub>2</sub>
$40 \times 20 + 257$	G5/P15	$71.1 \pm 3.2$	$70.0 \pm 3.2$	$85.2 \pm 2.4$	$81.1 \pm 2.7$	$81.5 \pm 2.0$	<b><math>88.8 \pm 2.7</math></b>
	G10/P10	$82.8 \pm 4.6$	$81.6 \pm 4.6$	$92.2 \pm 2.8$	$89.6 \pm 3.6$	$89.9 \pm 2.7$	<b><math>94.1 \pm 2.1</math></b>
$80 \times 20 + 257$	G5/P15	$72.3 \pm 4.1$	$71.4 \pm 4.1$	$85.4 \pm 2.9$	$83.7 \pm 4.2$	$82.9 \pm 3.3$	<b><math>90.8 \pm 2.1</math></b>
	G10/P10	$82.6 \pm 3.2$	$81.6 \pm 3.2$	$91.4 \pm 3.2$	$90.4 \pm 3.2$	$90.1 \pm 2.1$	<b><math>94.9 \pm 2.9</math></b>

Fig. 6 plotted the energies (in terms of  $l_2$  norm) for the columns of  $\mathbf{E}$ . One can see that the values of non-facial samples (last 257 columns in  $\mathbf{E}$ ) are obviously larger than that of facial samples. Therefore, the error term  $\mathbf{E}$  can also be used to detect the non-facial outliers. Namely the  $i$ -th sample in  $\mathbf{X}$  is considered as outlier if and only if  $\|[\mathbf{E}]_i\|_2 \geq \gamma$ . By setting the parameter  $\gamma = 2.2$ , the outlier detection accuracies<sup>5</sup> are 98.68% on the  $40 \times 20 + 257$  data and 99.19% on  $80 \times 20 + 257$  data, respectively.

<sup>5</sup>These accuracies are obtained by computing the percentage of correctly identified outliers. One may also consider the receiver operator characteristic (ROC) and compute its area under curve (AUC) [14] to evaluate the performance.

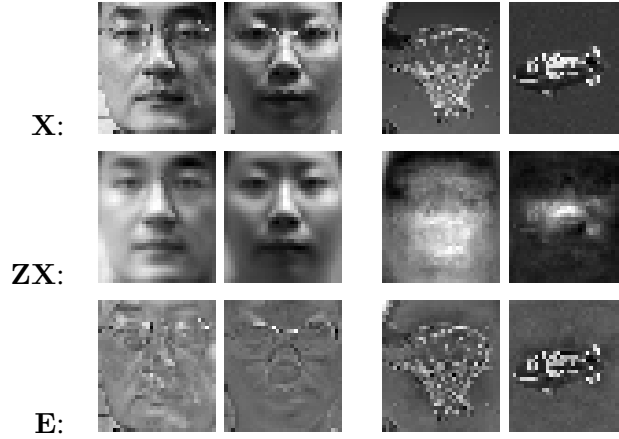


Fig. 5. Some examples of using TFRR to recover the intrinsic facial features and remove small errors and outliers (modeled by  $\mathbf{X} = \mathbf{Z}\mathbf{X} + \mathbf{E}$ ). The left two columns correspond to facial samples and the right two are non-facial samples. The middle row shows the features extracted by our algorithm ( $\mathbf{Z}\mathbf{X}$ ) and the bottom row shows the corruptions ( $\mathbf{E}$ ).

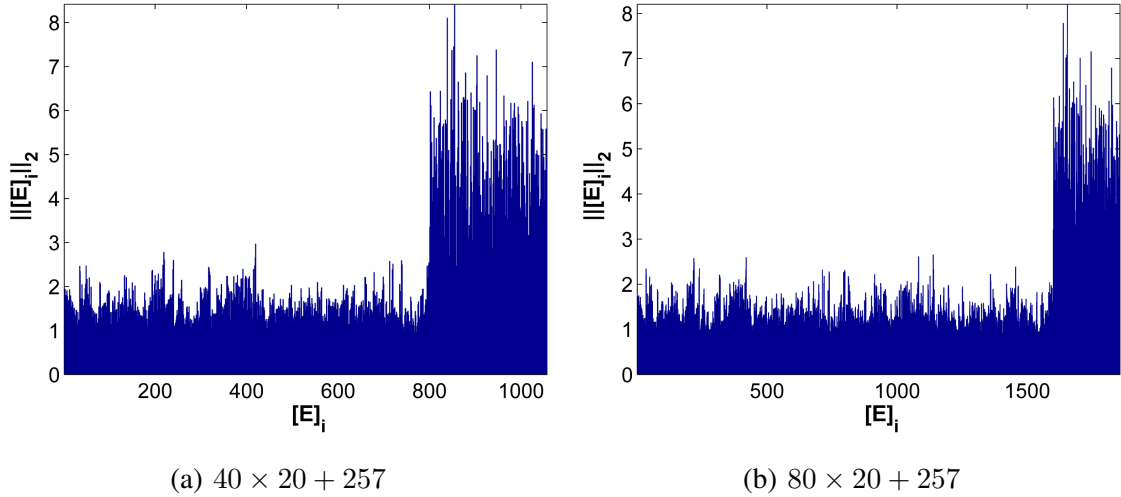


Fig. 6. The  $l_2$  norm for the columns of  $\mathbf{E}$ . The first 800 (a) and 1600 (b) columns are facial images and the last 257 columns are outliers.

## VII. CONCLUSIONS

This paper proposed a novel framework, named fixed-rank representation (FRR), for robust unsupervised visual learning. We proved that FRR can reveal the multiple subspace structure for clustering, even with insufficient observations. We also demonstrated that the transposed FRR (TFRR) can successfully recover the column space, and thus can be applied for feature extraction. There remain several directions for future work: 1) provide a deeper analysis on  $\mathbf{LR}$  (e.g., the general strategy for choosing efficient basis from  $\mathcal{R}(\mathbf{Z})$  for subspace clustering and determining dimension for feature extraction), 2) apply FRR to supervised and semi-supervised learning.

## ACKNOWLEDGMENT

This work is supported by the NSFC-Guangdong Joint Fund (No.U0935004), the NSFC Fund (No.61173103) and the Fundamental Research Funds for the Central Universities. R. Liu would also like to thank the support from CSC.

## REFERENCES

- [1] E. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” *Journal of the ACM*, vol. 58, no. 1, pp. 1–37, 2011.
- [2] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Trans. on PAMI*, vol. 31, no. 2, pp. 210–227, 2009.
- [3] R. S. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino, “Matrix completion for multi-label image classification,” in *NIPS*, 2011.
- [4] S. Rao, R. Tron, R. Vidal, and Y. Ma, “Motion segmentation in the presence of outlying, incomplete, and corrupted trajectories,” *IEEE Trans. on PAMI*, vol. 32, no. 10, pp. 1832–1845, 2010.
- [5] E. Elhamifar and R. Vidal, “Sparse subspace clustering,” in *CVPR*, 2009.
- [6] G. Liu, Z. Lin, and Y. Yu, “Robust subspace segmentation by low-rank representation,” in *ICML*, 2010.
- [7] P. Favaro, R. Vidal, and A. Ravichandran, “A closed form solution to robust subspace estimation and clustering,” in *CVPR*, 2011.
- [8] B. Recht, M. Fazel, and P. Parrilo, “Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization,” *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.
- [9] E. Elhamifar and R. Vidal, “Clustering disjoint subspaces via sparse representation,” in *ICASSP*, 2010.
- [10] M. Soltanolkotabi and E. Candès, “A geometric analysis of subspace clustering with outliers,” *Technical Report (arXiv:1112.4258)*, 2011.
- [11] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. on PAMI*, vol. 22, no. 8, pp. 888–905, 2000.
- [12] B. Nasihatkon and R. Hartley, “Graph connectivity in sparse subspace clustering,” in *CVPR*, 2011.
- [13] Y. Ni, J. Sun, X. Yuan, S. Yan, and L. Cheong, “Robust low-rank subspace segmentation with semidefinite guarantees,” in *ICDM Workshop*, 2010.
- [14] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, “Robust recovery of subspace structures by low-rank representation,” *submitted to IEEE Tran. on PAMI*, 2011.
- [15] J. Costeira and T. Kanade, “A multibody factorization method for independently moving objects,” *IJCV*, vol. 29, no. 3, pp. 159–179, 1998.
- [16] G. Liu and S. Yan, “Latent low-rank representation for subspace segmentation and feature extraction,” in *ICCV*, 2011.
- [17] Z. Lin, M. Chen, L. Wu, and Y. Ma, “The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices,” *UIUC Technical Report UILU-ENG-09-2215*, 2009.
- [18] Z. Lin, R. Liu, and Z. Su, “Linearized alternating direction method with adaptive penalty for low rank representation,” in *NIPS*, 2011.
- [19] A. Fisher, “The statistical utilization of multiple measurements,” *Annals of Eugenics*, vol. 8, pp. 376–386, 1938.
- [20] H. Hotelling, “Relations between two sets of variates,” *Biometrika*, vol. 28, pp. 321–377, 1936.
- [21] F. De la Torre, “A least-squares framework for component analysis,” *IEEE Trans. on PAMI*, vol. 34, no. 6, pp. 1041–1055, June 2012.
- [22] K. Pearson, “On lines and planes of closest fit to systems of points in space,” *The London, Edinburgh and Dublin Philosophical Magazine and Journal*, vol. 6, pp. 559–572, 1901.

- [23] H. Hotelling, “Analysis of a complex of statistical variables into principal components,” *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933.
- [24] J. Liu, S. Ji, and J. Ye, “Multi-task feature learning via efficient  $l_{2,1}$ -norm minimization,” in *UAI*, 2009.
- [25] G. Golub and C. Van Loan, *Matrix Computations*, 3rd ed. Johns Hopkins University Press, 1996.
- [26] C. Eckart and G. Young, “The approximation of one matrix by another of lower rank,” *Psychometrika*, 1936.
- [27] G. Liu, H. Xu, and S. Yan, “Exact subspace segmentation and outlier detection by low-rank representation,” *submitted to JMLR (arXiv:1109.1646)*, 2011.
- [28] Y. Shen, Z. Wen, and Y. Zhang, “Augmented Lagrangian alternating direction method for matrix separation based on low-rank factorization,” *preprint*, 2011.
- [29] M. Fischler and R. Bolles, “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [30] J. Yan and M. Pollefeys, “A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and nondegenerate,” in *ECCV*, 2006.
- [31] R. Tron and R. Vidal, “A benchmark for the comparison of 3D motion segmentation algorithms,” in *CVPR*, 2007.
- [32] P. Philips, P. Flynn, T. Scruggs, and K. Bowyer, “Overview of the face recognition grand challenge,” in *CVPR*, 2005.
- [33] G. Griffin, A. Holub, and P. Perona, “Caltech-256 object category dataset,” *Caltech Technical Report*, 2007.
- [34] X. He and P. Niyogi, “Locality preserving projections,” in *NIPS*, 2003.
- [35] X. He, D. Cai, S. Yan, and H. Zhang, “Neighborhood preserving embedding,” in *ICCV*, 2005.